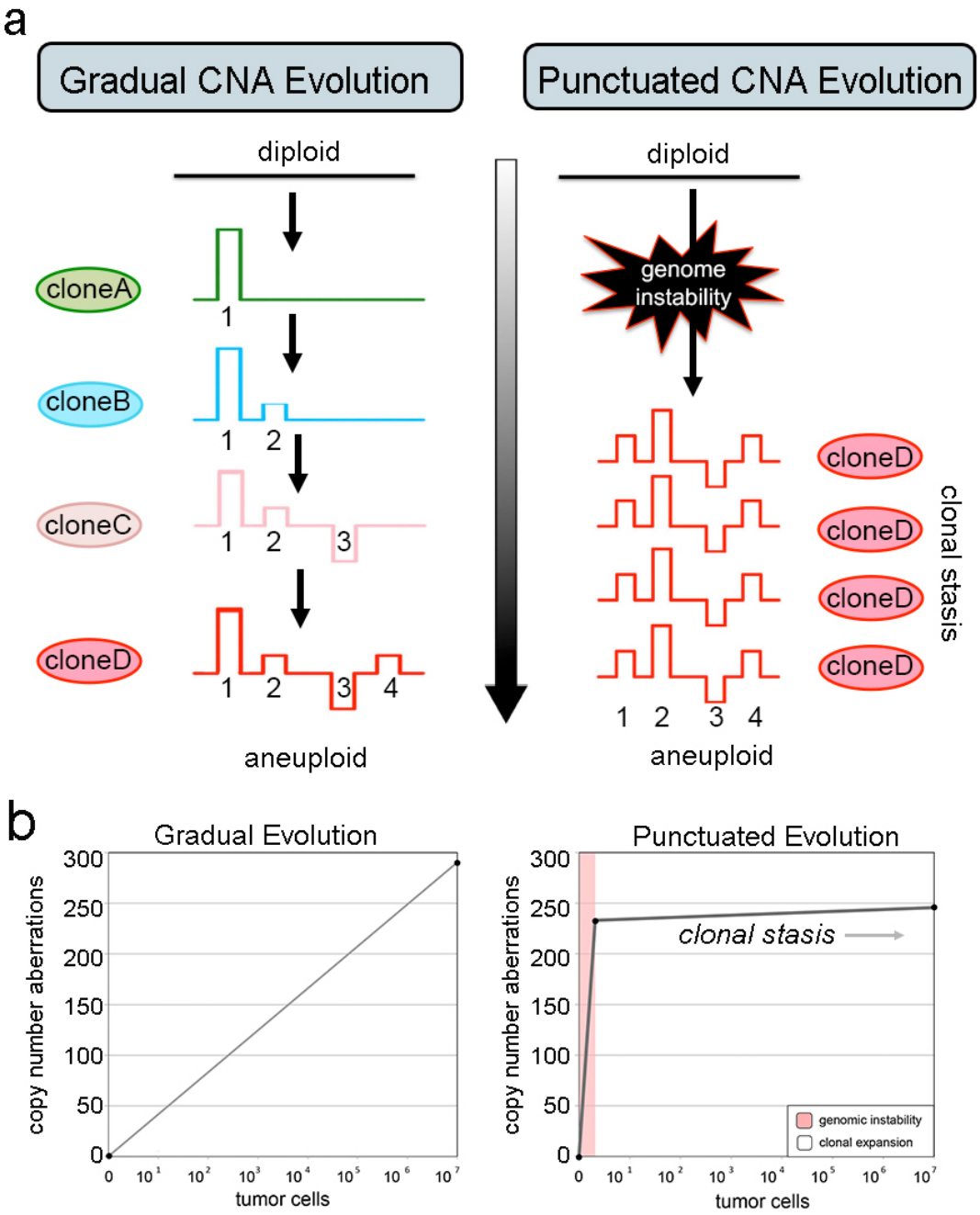
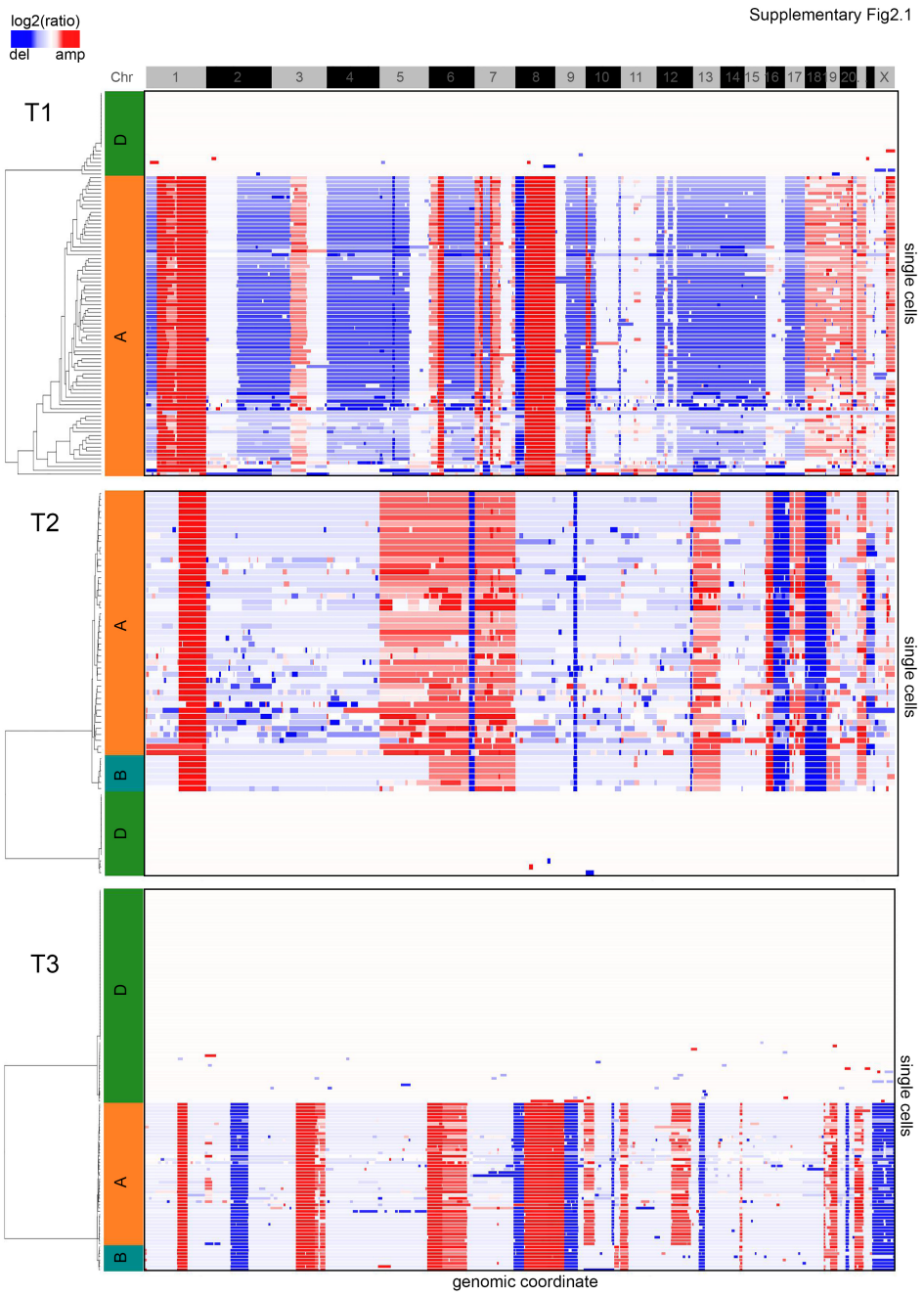
# Punctuated Copy Number Evolution and Clonal Stasis in Triple-Negative Breast Cancer

**Gao et al. 2016**

**CONTENTS**

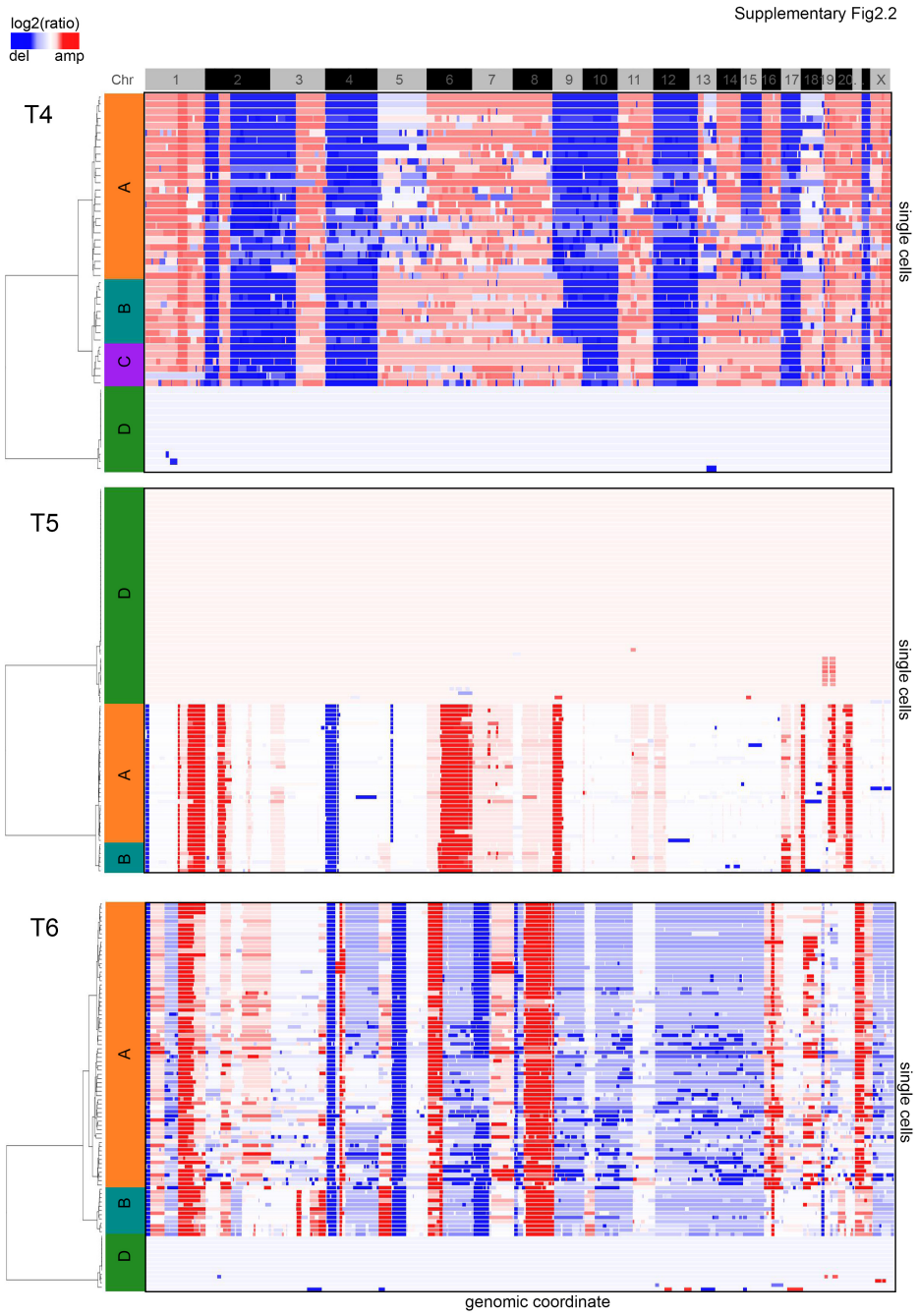



**Supplementary Figure 1 - Models of Copy Number Evolution in Tumors**
**(a)** Gradual model of copy number evolution. Diploid cells acquire CNAs sequentially over extended periods of time, leading to more malignant stages of cancer. This model predicts a large number of intermediate clones (A, B, C) that evolve during the transition from diploid to aneuploid genomes. **(b)** Punctuated model of copy number evolution. Diploid cells undergo a punctuated burst of evolution in which many chromosomal amplifications and deletions are acquired in a short period of time at the earliest stages of tumor progression. Genome instability is then turned off, and the clones undergo stable clonal expansions of aneuploid profiles during tumor growth to form the tumor mass.



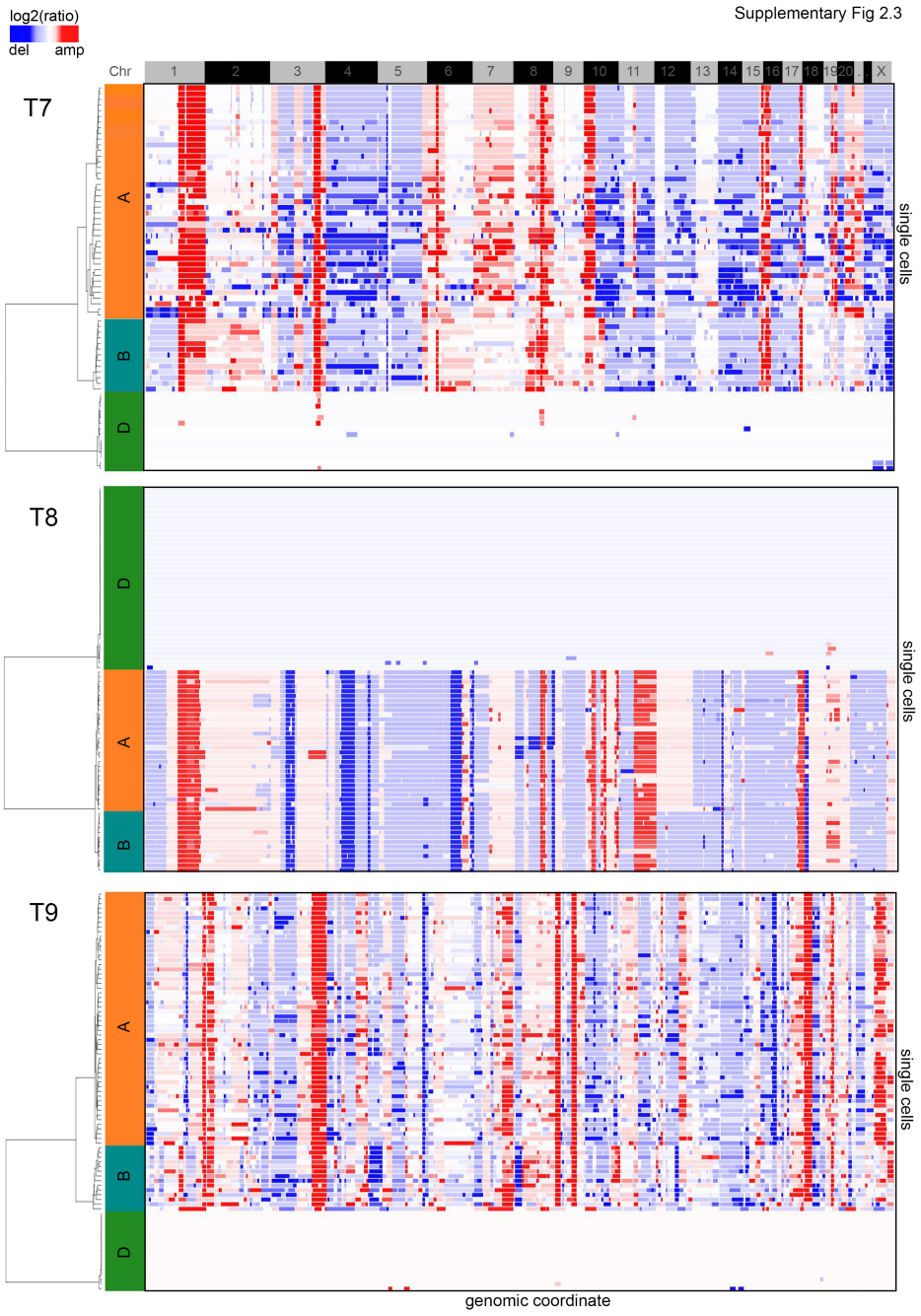

**Supplementary Figure 2 - Clustered Heatmaps of Diploid and Aneuploid Single Cell Copy Number Profiles.** One-dimensional hierarchical clustering was performed on the single cell copy number profiles from all diploid and aneuploid cells from each TNBC patient. Copy number aberrations are plotted on the X-axis in genome order, while single cells are plotted on the Y-axis. Clonal subpopulations are color coded by row sidebars as follows: diploid cells (D) and tumor subpopulations (A, B, and C).

**ST2.1** – Clustered heatmaps of patients T1, T2 and T3



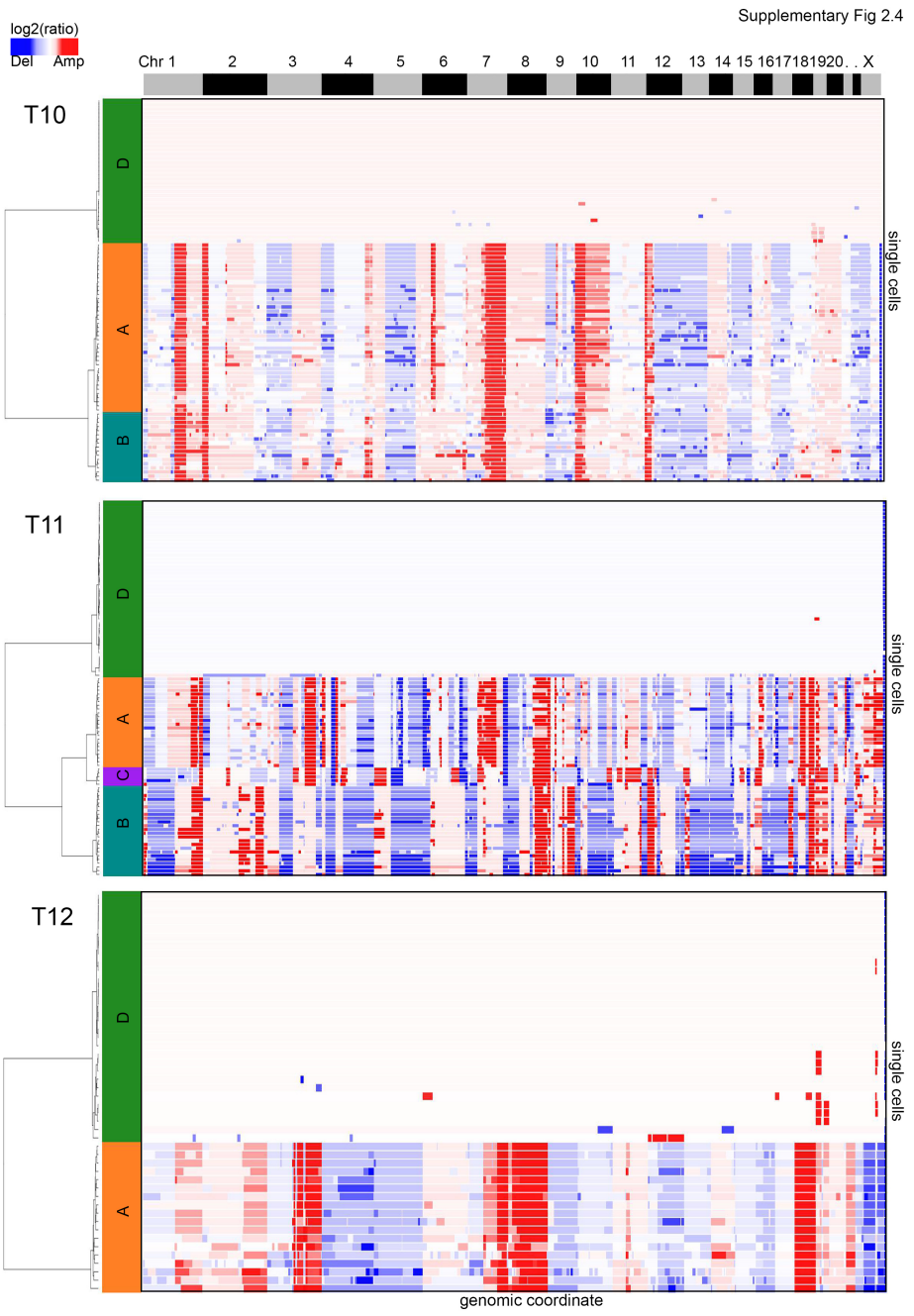

**Supplementary Figure 2 – continued**

**ST2.2** – Clustered heatmaps of patients T4, T5 and T6



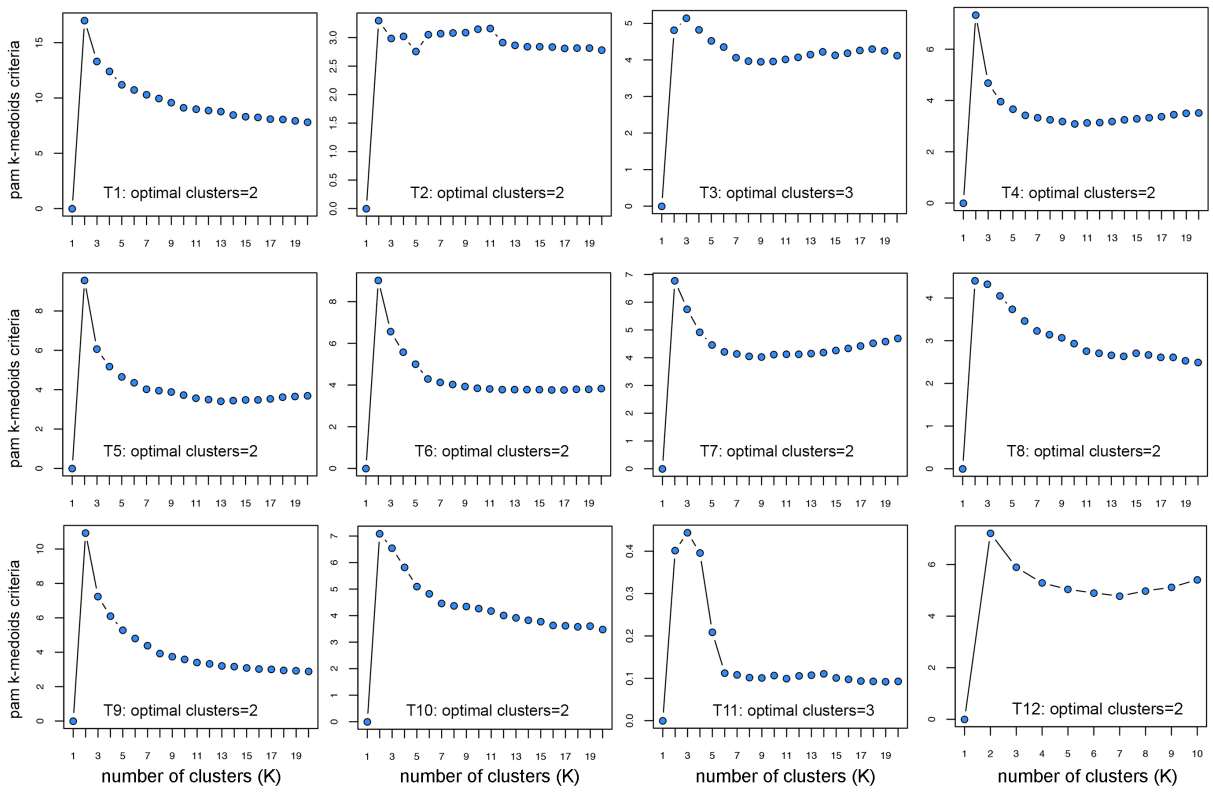

**Supplementary Figure 2 – continued**

**ST2.3** – Clustered heatmaps of patients T7, T8 and T9

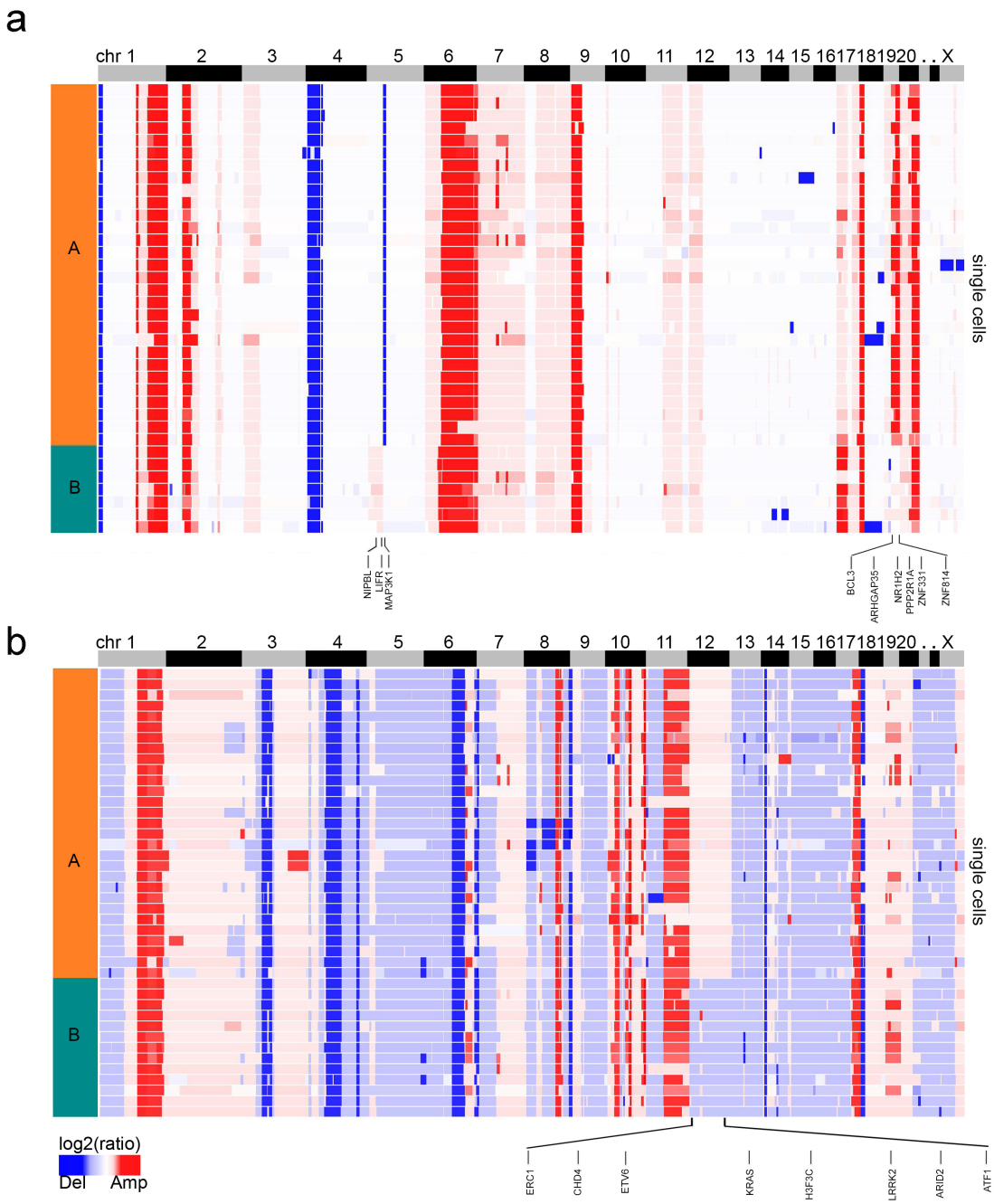



**Supplementary Figure 2 – continued**

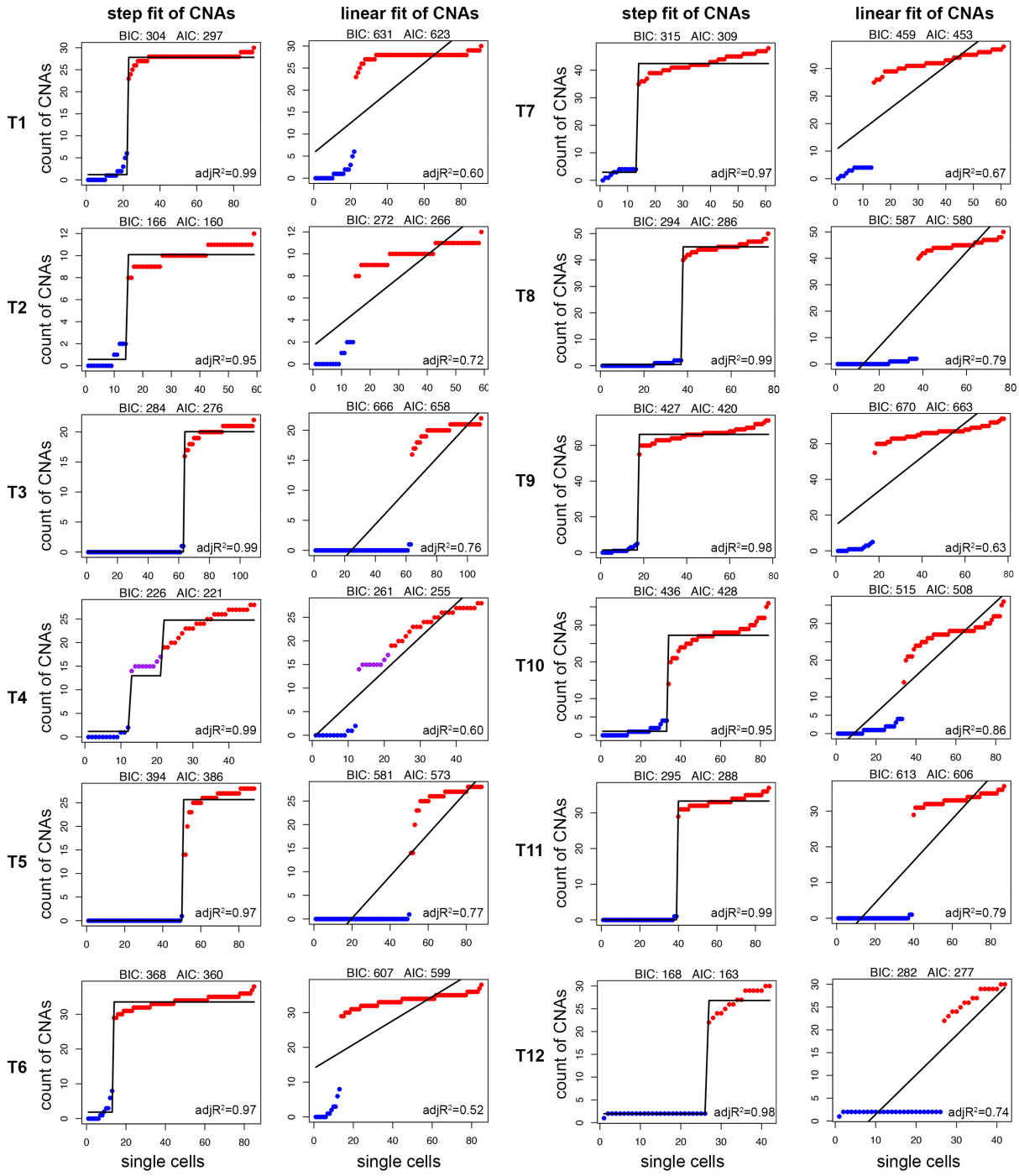
**ST2.4** – Clustered heatmaps of patients T10, T11 and T12



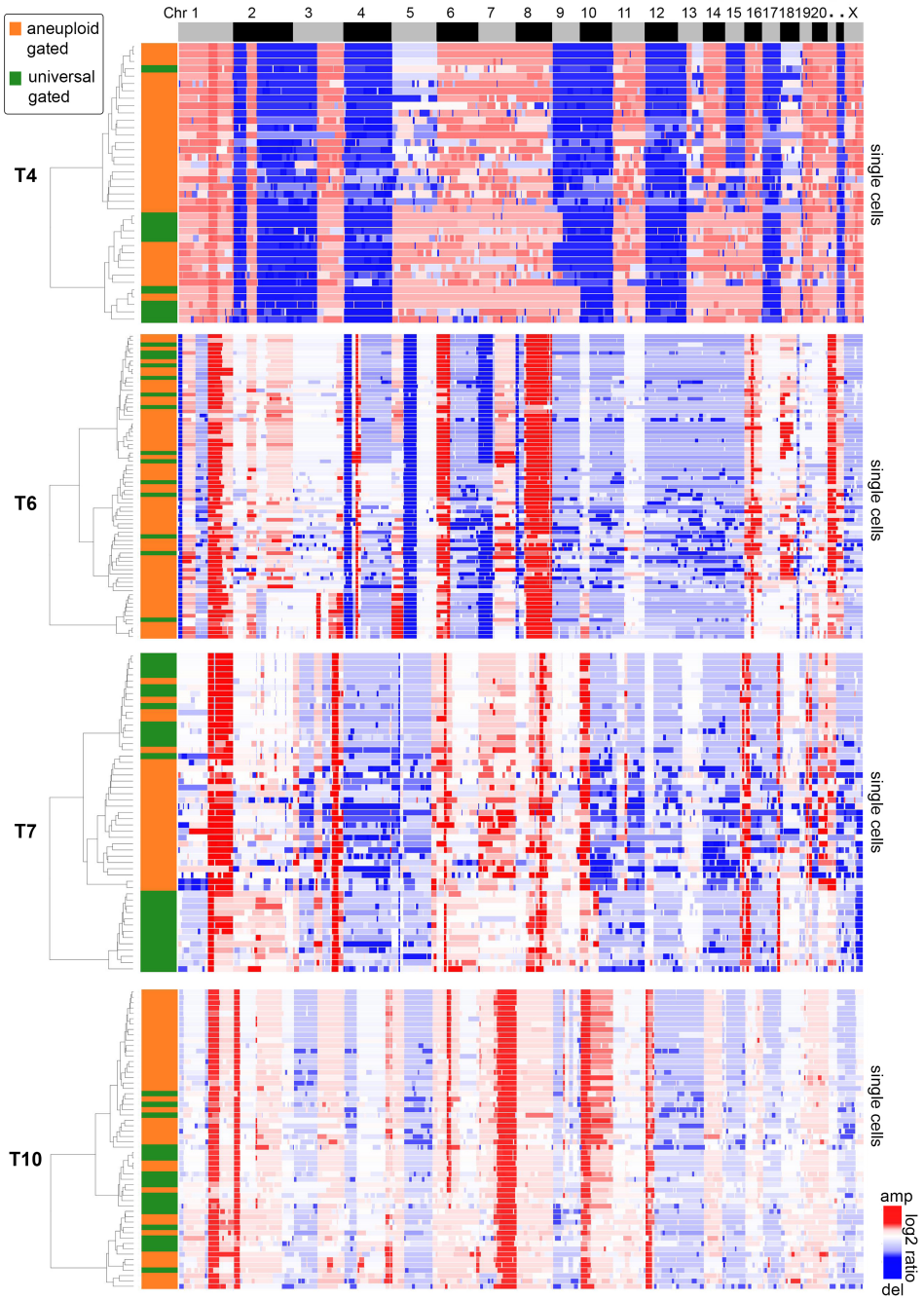

**Supplementary Figure 3 – Optimized Cluster Selection using PAMK Medoids**
Partitioning around medoids clustering of single cell aneuploid profiles for each of the 12 TNBC patient. The optimal number of clusters was selected by the highest peak using the Calinski-Harabasz index or optimum average silhouette width (T11). The K-range of 1 to 20 clusters was tested for each TNBC patient (X-axis).



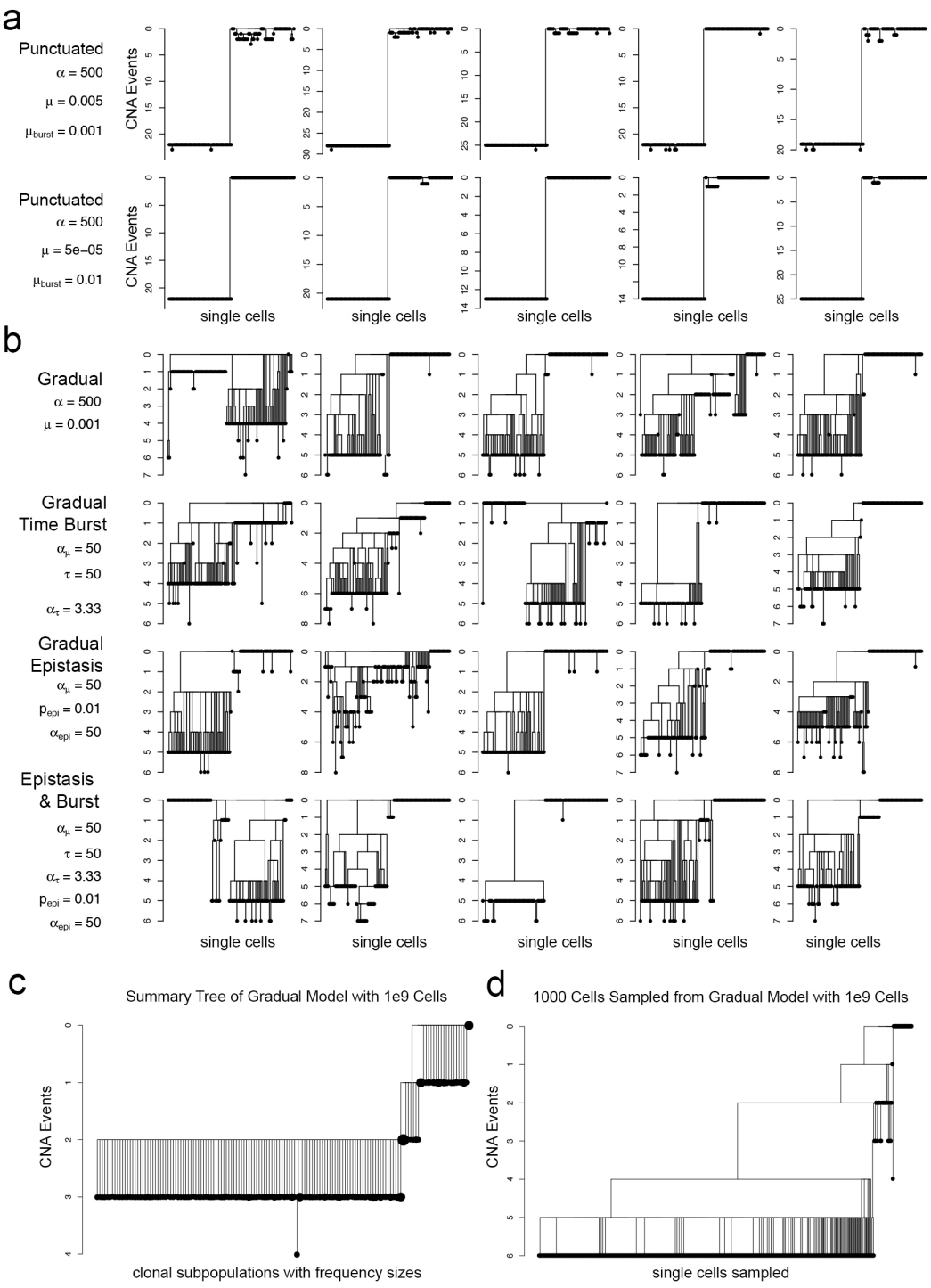

**Supplementary Figure 4 - Divergent Subpopulations in Polygenomic Tumors T5 and T8**
One-dimensional clustered heatmaps of aneuploid single cell copy number profiles from polygenomic tumors T5 (**a**) and T8 (**b**). The major subpopulations (A, B) are color coded by row sidebars. Copy number profiles of single cell are arranged in genome order. Cancer genes in CNA regions that distinguish the subpopulations are annotated.



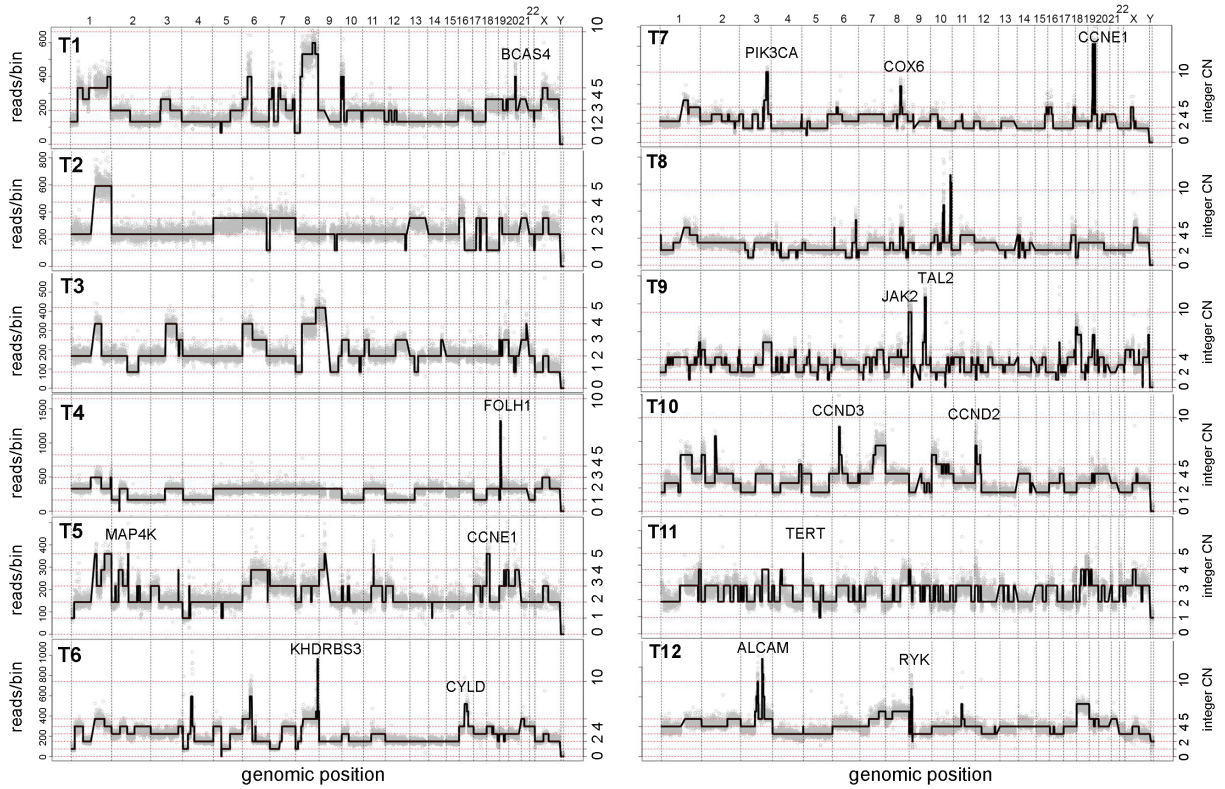

**Supplementary Figure 5 - Step and Slope Fitting of Sorted Single Cell CNA Count Data**
The number of segments in each single cell was counted for each tumor and the data was sorted by CNA number. Step functions or linear regression were fitted to the data. To determine better fit, three metrics were calculated: Bayesian information criteria (BIC), the Akaike information criteria (AIC) values, and adjusted R-squares values.



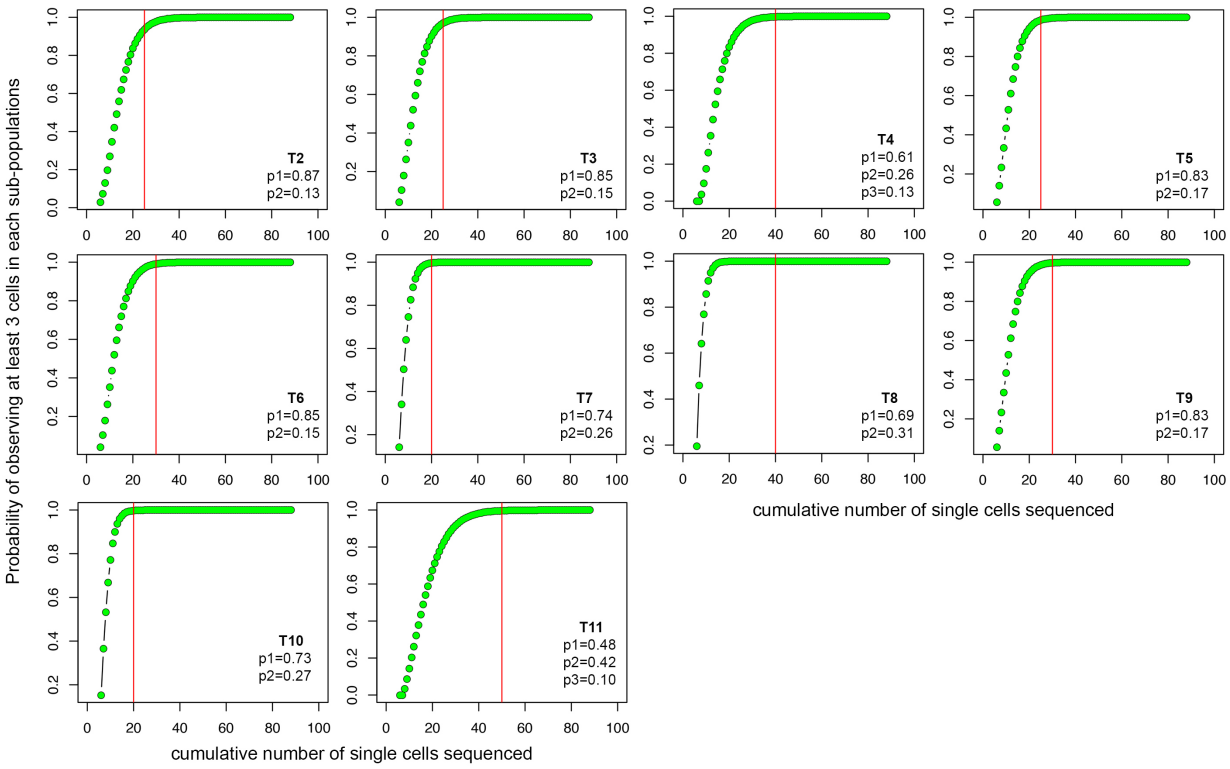

**Supplementary Figure 6 – Universal and Gated Single Cell CNA Profiles**
Clustered heatmaps of single tumor cells flow-sorted from aneuploid fractions (orange) and ungated fractions (green) for four TNBC patients: T4, T6, T7 and T10. Diploid cells from ungated fractions or gated populations were removed from this analysis. Copy number profiles are arranged in genome order on the X-axis, while single cells are plotted along the Y-axis.




**Supplementary Figure 7 – Trees Constructed from Gradual and Punctuated Evolution Modeling Data.** (**a**) Trees constructed by randomly sampling 100 single cells from punctuated burst modeling results (100 million cells) with different fitness and burst mutation rates. (**b**) Trees of 100 single cells sampled from gradual modeling results (100 million cells) with different simulation conditions for epistatic interactions, burst timing and combinations thereof. See supplementary notes for more details.




**Supplementary Figure 8 - Focal Amplifications in TNBCs.**
Consensus copy number profiles representing the population of tumor cells were calculated for each tumor using the single cell aneuploidy profiles. High copy number focal amplifications with cancer genes are annotated.




**Supplementary Figure 9 – Single Cell Sampling Sizes and Detection Power**
Posterior saturation curves of single cell sample sizes constructed for each polygenomic TNBC tumor using the experimentally measured subpopulation frequencies. The posterior probability of observing at least 3 single cells in each subpopulation given number of cells sequenced was calculated based on a binomial distribution for the 8 biclonal tumors and multinomial distribution for two triclonal tumors (T4 and T11). Saturation was expected to occur by sequencing 20-40 cells in all tumors. Monoclonal tumors (T1 and T12) were excluded from the analysis.

## Supplementary Table 1 - Clinical Information for 12 TNBC Tumors

| Tumor ID | Tumor Size | Tumor | Grade | Adjuvant Chemotherapy | Age at Diagnosis | Lymph Nodes | Recurrence | Metastasis - Local/Regional | Metastasis - Distant | Sub Populations | Diversity Index |
|---|---|---|---|---|---|---|---|---|---|---|---|
| T1 | 1.9 cm | Invasive ductal | III | paclitaxel, FAC | 35 | Yes | Yes | Axillary Lymph Nodes | Brain | 1 | 0 |
| T2 | 6.0 cm | Invasive ductal | III | Taxotere, gemcitabine, zometa | 66 | Yes | Yes | Internal Mammary & Axillary Lymph Nodes | Bone | 2 | 0.381 |
| T3 | 1.6 cm | Invasive ductal | III | paclitaxel, AC | 69 | No | No | No | No | 2 | 0.5 |
| T4 | 2.7 cm | Invasive ductal | III | paclitaxel, FAC | 65 | No | No | No | No | 3 | 0.922 |
| T5 | 1.6 cm | Invasive ductal | III | paclitaxel, FAC | 54 | Yes | Yes | Supraclavicular Lymph Node | Brain | 2 | 0.485 |
| T6 | 1.8 cm | Invasive ductal | III | docetaxel, cyclophosphamide | 69 | No | No | No | No | 2 | 0.423 |
| T7 | 2.5 cm | Invasive ductal | III | paclitaxel, AC | 79 | No | No | No | No | 2 | 0.567 |
| T8 | 1.3 cm | Invasive ductal | III | not known | 53 | No | No | No | No | 2 | 0.618 |
| T9 | 6 cm | Invasive ductal | III | paclitaxel, FEC | 57 | Yes | No | Axillary Lymph Nodes | No | 2 | 0.463 |
| T10 | 1.4 cm | Invasive ductal | III | paclitaxel, FEC | 37 | No | No | No | No | 2 | 0.581 |
| T11 | 3.2 cm | Invasive ductal | III | paclitaxel, AC | 63 | Yes | No | na | no | 3 | 0.952 |
| T12 | 2.4cm | Invasive ductal | III | paclitaxel, AC | 46 | No | No | na | no | 1 | 0 |

This table lists the clinical information for the 12 invasive ductal carcinoma samples that were analyzed in this study. All samples are high grade (III) and were surgically resected prior to administration of adjuvant chemotherapy. Chemotherapy abbreviations are: AC (doxorubicin and cyclophosphamide), FAC (5FU, doxorubicin and cyclophosphamide) and FEC (epirubicin and cyclophosphamide). Columns listed in the table include: tumor ID; tumor size; classification; grade; adjuvant chemotherapy; age at diagnosis; lymph node metastases; recurrence; local or regional metastasis; distant metastasis; Shannon diversity index; number of major clonal subpopulations.

**Supplementary Table 2 - Sequencing Metrics for Single Cell Data**

| Tumor | Number of cells sequenced | Average reads per bin | Average reads per cell | Average DNA ploidy |
|---|---|---|---|---|
| T1 | 100 | 206 | 2,517,807 | 2.0N, 3.1N |
| T2 | 65 | 260 | 3,172,745 | 2.0N, 2.2N |
| T3 | 120 | 184 | 2,250,271 | 2.0N, 2.4N |
| T4 | 54 | 326 | 3,982,240 | 2.0N, 1.8N, 3.6N |
| T5 | 90 | 171 | 2,090,624 | 2.0N, 2.7N |
| T6 | 92 | 215 | 2,623,868 | 2.0,N 2.9N |
| T7 | 68 | 114 | 1,395,184 | 2.0N, 3.1N |
| T8 | 84 | 123 | 1,503,512 | 2.0N, 2.7N |
| T9 | 85 | 179 | 2,190,243 | 2.0N, 3.3N |
| T10 | 94 | 146 | 1,787,844 | 2.0N, 3.3N |
| T11 | 100 | 537 | 6,549,081 | 2.0N, 1.7N, 2.9N, 3.1N |
| T12 | 48 | 887 | 10,828,886 | 2.0N, 4.1N |

This table lists the next-generation sequencing metrics for the single cell copy number data for each TNBC patient. Columns include: tumor ID; number of cells sequenced; average number of reads per 220kb bin; average number of sequencing reads for each single cell; mean DNA ploidy distributions detected in the tumor.

## Supplementary Table 3 – Clonal Frequencies of Tumor Subpopulations

| Tumor | Total Cells Sequenced | Clonal Aneuploid cells | Number of Subclonal Tumor cells | | | Clonal Frequency | | |
|---|---|---|---|---|---|---|---|---|
| | | | cloneA | cloneB | cloneC | cloneA | cloneB | cloneC |
| T1 | 100 | 72 | 72 | NA | NA | 1.00 | NA | NA |
| T2 | 65 | 47 | 41 | 6 | NA | 0.87 | 0.13 | NA |
| T3 | 120 | 47 | 40 | 7 | NA | 0.85 | 0.15 | NA |
| T4 | 54 | 38 | 23 | 10 | 5 | 0.61 | 0.26 | 0.13 |
| T5 | 90 | 36 | 30 | 6 | NA | 0.83 | 0.17 | NA |
| T6 | 92 | 73 | 62 | 11 | NA | 0.85 | 0.15 | NA |
| T7 | 68 | 50 | 37 | 13 | NA | 0.74 | 0.26 | NA |
| T8 | 84 | 42 | 29 | 13 | NA | 0.69 | 0.31 | NA |
| T9 | 85 | 63 | 52 | 11 | NA | 0.83 | 0.17 | NA |
| T10 | 94 | 56 | 41 | 15 | NA | 0.73 | 0.27 | NA |
| T11 | 100 | 48 | 23 | 20 | 5 | 0.48 | 0.42 | 0.10 |
| T12 | 48 | 16 | 16 | NA | NA | 1.00 | NA | NA |

This table lists the total number of tumor cells sequenced and number of tumor cells that were clustered into designated clonal subpopulation. Clonal frequencies were calculated from the genotype frequencies of the subpopulations in each TNBC patient (online methods).

## Supplementary Table 4 – Non-clonal Cell Frequencies in TNBC Tumors

| Tumor | Total Cells Sequenced | Clonal Aneuploid Cells | Diploid Cells | Metastable Aneuploid Cells | Metastable Cells (%) | Pseudodiploid Cells | Pseudodiploid (%) | Chromazemic Cells | Chromazemic cells (%) |
|---|---|---|---|---|---|---|---|---|---|
| T1 | 100 | 72 | 19 | 6 | 8.33 | 3 | 15.79 | 0 | 0.00 |
| T2 | 65 | 47 | 13 | 3 | 6.38 | 1 | 7.69 | 1 | 1.54 |
| T3 | 120 | 47 | 61 | 6 | 12.77 | 5 | 8.20 | 1 | 0.83 |
| T4 | 54 | 38 | 11 | 3 | 7.89 | 1 | 9.09 | 1 | 1.85 |
| T5 | 90 | 36 | 47 | 3 | 8.33 | 3 | 6.38 | 1 | 1.11 |
| T6 | 92 | 73 | 11 | 6 | 8.22 | 2 | 18.18 | 0 | 0.00 |
| T7 | 68 | 50 | 11 | 4 | 8.00 | 2 | 18.18 | 1 | 1.47 |
| T8 | 84 | 42 | 38 | 1 | 2.38 | 1 | 2.63 | 1 | 1.19 |
| T9 | 85 | 63 | 16 | 5 | 7.94 | 1 | 6.25 | 0 | 0.00 |
| T10 | 94 | 56 | 34 | 2 | 3.57 | 1 | 2.94 | 1 | 1.06 |
| T11 | 100 | 48 | 46 | 5 | 10.42 | 1 | 2.17 | 0 | 0.00 |
| T12 | 48 | 16 | 28 | 2 | 12.50 | 2 | 7.14 | 0 | 0.00 |

This table lists the total number of clonal aneuploid, diploid and non-clonal cells detected in each tumor. Nonclonal cells are classified as metastable aneuploid cells, pseudodiploid cells and chromazemic cells based on their copy number profiles. The percentages of nonclonal cells in each tumor are also listed.

## Supplementary Table 5 – Distribution of CNAs in Pseudodiploid Cells

| PD cells | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | | Loss Count | Loss Freq | Loss Odds |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| chr1 | | | | | | | | | | | | | | | | | | | | | | | | 0 | 0.00 | 0.00 |
| chr2 | | | | | | | | | | | | | | | | | | | | -1 | | | | 1 | 0.05 | 0.05 |
| chr3 | | | | | | -1 | | | | | | | | | | | | | | | | | | 1 | 0.05 | 0.05 |
| chr4 | | | | | | | | | | | | | | | | | | | | | | | | 0 | 0.00 | 0.00 |
| chr5 | | 1 | -1 | | | | | | | | | | -1 | | | | | | | | | | | 2 | 0.09 | 0.10 |
| chr6 | | | | | | | | | -1 | -1 | | | | | | | | | | | | | | 2 | 0.09 | 0.10 |
| chr7 | | | | | | | | | | | | | | | | | | | | | | | | 0 | 0.00 | 0.00 |
| chr8 | | | | | | | | | | | | | | | | -1 | | | | | 1 | -1 | | 2 | 0.09 | 0.10 |
| chr9 | | | | | | | | | | | | | | | | | | | | | 1 | | | 0 | 0.00 | 0.00 |
| chr10 | 1 | | | | | | | | | | | | | | | | | | -1 | | | | | 1 | 0.05 | 0.05 |
| chr11 | | | | | | | | | | | | | | | | | | | | | | | | 0 | 0.00 | 0.00 |
| chr12 | | | | | | | | | | | | | | | | | | 1 | | | | | | 0 | 0.00 | 0.00 |
| chr13 | | | | | | -1 | | | | | -1 | | | | | | | | | | | | | 2 | 0.09 | 0.10 |
| chr14 | | -1 | | | | | | | | | | | | | | | | | -1 | | | | | 2 | 0.09 | 0.10 |
| chr15 | | | | | -1 | | | | | | | | | | | | | | | | | | | 1 | 0.05 | 0.05 |
| chr16 | | | | | | | -1 | | | | | | | | | | | | | | | | | 1 | 0.05 | 0.05 |
| chr17 | | | | | | 1 | | | | | | | | | | | | | | | | | | 0 | 0.00 | 0.00 |
| chr18 | | | | | | | | | | | | | | | | | | | | | | | | 0 | 0.00 | 0.00 |
| chr19 | | | | | | | | | | | | | | | | | -1 | | | | | | | 1 | 0.05 | 0.05 |
| chr20 | | | | | | | | | | | | | | | | | | | | | | | | 0 | 0.00 | 0.00 |
| chr21 | | | | | | | | | | | | | | | | | | | | | | | | 0 | 0.00 | 0.00 |
| chr22 | | | | | | | | | | | | | | | | | | | | | | | | 0 | 0.00 | 0.00 |
| chrX | | | | -1 | | | | -1 | | | | -1 | | -1 | -1 | | | | | | | | | 5 | 0.23 | 0.29 |

This table shows the genomic distribution of copy number amplifications (red) and deletions (blue) detected in the pseudodiploid (PD) cells identified in all TNBC patients. This plot shows a statically significant enrichment of X chromosome loss.

**Supplementary Table 6 –Pseudodiploid Cells in Matched Normal Breast Tissues**

| Matched Normal Tissue | Total Cells Sequenced | Diploid Cells | Pseudodiploid Cells | Diploid(%) | Pseudodiploid(%) |
|---|---|---|---|---|---|
| T3N | 32 | 30 | 2 | 93.8 | 6.3 |
| T5N | 38 | 36 | 1 | 94.7 | 5.3 |
| T8N | 25 | 26 | 1 | 96.3 | 3.7 |
| T10N | 24 | 22 | 2 | 91.7 | 8.3 |

This table lists the percentages and cell counts of pseudo-diploid cells in matched normal breast tissues from 4 patients.

## Supplementary Notes on Gradual and Punctuated Tumor Modeling

To investigate alternative hypotheses for the generation of the patient tumor phylogenies, we designed a multi-type birth-death-mutation stochastic process model. The model is initiated with a single ancestor cancer cell, which divides and dies at rates $b_0$ and $d_0$ per day, respectively. Upon a cell division event, one of the daughter cells may accumulate a new copy number alteration (CNA) with probability $u_0$ and thus initiates a new subclone. This subclone has new birth rate:

$$b_1 = b_0 + s$$

Where s is chosen from a mutational fitness distribution. Newly emerging CNAs are considered unique and not previously observed in the population, and are irreversible. This branching process model forms the basis of all scenarios investigated. Individual scenarios differ by utilizing alternative assumptions regarding parameters such as epistatic fitness effects or mutation rates (see below). The stochastic process continues until the population reaches size N, which is chosen based on the patient tumor volume at the time of sampling. We implemented this model as an exact stochastic computer simulation. Within each simulation, we designated cells as being either "diploid" or "aneuploid" based on a cutoff of 5 CNAs per tumor cell genome. This cutoff was chosen because we considered the difference between diploid and aneuploid cells to be greater than 5 CNAs. Once a simulation run reaches N cells, we obtained a random sample of 50 cells from the aneuploid population; this number was chosen since on average, we sequenced 50 single aneuploidy tumor cells per patient. For each sample, we constructed the Hamming distance matrix of the 50 aneuploid cells with 50 ancestor diploid cells and use their labels to run AMOVA[24], an algorithm designed to test whether the genetic diversity between two groups (in our case, aneuploid and ancestral cells) is significantly different. Since AMOVA is based on the ratio of squared distances between groups to that within groups, aneuploid cells are required to have a common ancestor in order to observe a large F-statistic. We performed permutation tests to determine the significance of the F-statistic by rerunning AMOVA with labels (aneuploid and dipoid) randomly assigned multiple times; this approach allowed us to create a distribution for the test statistic that we could then compare our F-statistic to determine how likely it is that we would observe our sample if the underlying tree did not have a distinct subclone.

We simulated 50 independent runs of the model for each parameter combination, and sampled with replacement 100 cells from each resulting branching process. In order to determine parameters that were able to recapitulate the phylogenies observed from patients, we created rules for each sample based on a p-value and the number of subclones present, and for each simulation based on the number of samples, adhering to the following criteria:

(i) Samples must have a p-value below 0.1 and a single branch leading to the aneuploid tumor cells (determined by sharing a minimum distance of 5 CNAs from the root)
(ii) 90% of samples taken from a single simulation must satisfy (i)
(iii) Parameters must have at least 90% of simulations that satisfy (ii).

We tested all scenarios at different parameter combinations for a total of 2,022 different models until the population hit $2\times10^6$ cells in order to determine likely scenarios or parameters at a reduced scale. Thereafter, we performed additional simulation runs across a range of parameters similar to the previous sets used until the number of cells reached $\approx 2.86\times10^7$ cells, which is equivalent to the total number of aneuploid tumor cells obtained from our patient data set (flow-sorting cell counts of the aneuploid fractions). Within these simulations, we refined our first criterion for selecting samples to account for observing a branch between diploid and aneuploid cells with a larger distance:

(i) Samples must have a p-value below 0.1 and a single branch for tumor cells (determined by sharing a minimum distance of 5 SCNAs from the root) and must contain a branch of at least length 5.

**Alternative Assumptions Utilized in Our Evolutionary Branching Process Models**

The following sections describe the underlying rules behind each of the alternative scenarios we considered along with the parameters determining each. The set of models we considered is obviously not exhaustive, and other possibilities may exist, but we attempted to include models that seemed biologically reasonable while allowing the branching process to grow as naturally as possible (i.e. minimizing the number of rules created to avoid overfitting).

All scenarios are based on a birth-death-mutation process initiating with a single ancestor cell. The growth of its clone is determined by the ancestor birth and death rates, which are held constant across individual scenarios for comparability. New mutations in the process occur with some probability and new mutations have a birth rate generated from the sum of the parent fitness and a random variable with a distribution discussed in the following sections. Thus, for a cell of type $i$ having birth rate $b_i$, death rate $d_i$, and mutation probability $u_i$, the elementary transition probabilities of the stochastic process are given by:

$$\begin{aligned} birth\colon P\big(X_i(t+h) = X_i(t)+1 \big| X_i(t)\big) &= b_i(1-u_i)h + o(h) \\ death\colon P\big(X_i(t+h) = X_i(t)-1 \big| X_i(t)\big) &= d_i h + o(h) \\ mutation\colon P\big(X_{i+1}(t+h) = 1 \big| X_i(t)\big) &= b_i u_i h + o(h) \end{aligned}$$

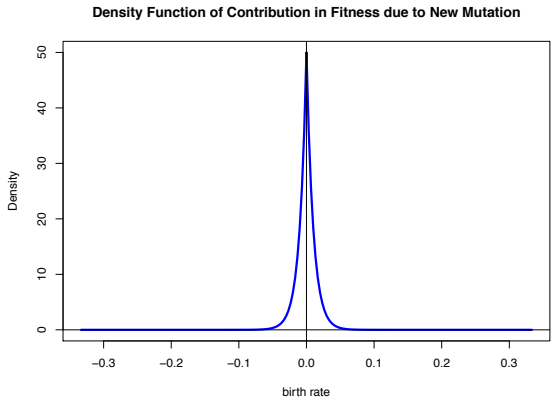
Thus, for a cell of type $i$ having birth rate $b_i$, death rate $d_i$, and mutation probability $u_i$, The time a cell dies or divides is exponentially distributed with rate $(b_i + d_i)$. At this time, the cell divides with probability $\frac{b_i(1-u_i)}{b_i+d_i}$, dies with probability $\frac{d_i}{b_i+d_i}$, or mutates into a new cell type with probability $\frac{b_i u_i}{b_i+d_i}$. After a mutation, a type $i+1$ cell appears. We determine the parameters for a type $i+1$ cell according to the following models.

**1. The gradual model**

To investigate gradual CNA accumulation using our model, we first considered an exponentially decreasing fitness distribution with a constant mutation probability over time. During each mutation event, a single CNA arises. If a parent has birth rate $b_0$ and gives rise to a mutant daughter, then the latter has birth rate $b_1 = b_0 + s$ where s is a random variable with

exponentially decreasing tails. The probability density function (p.d.f.) of s (Fig. S1) is:

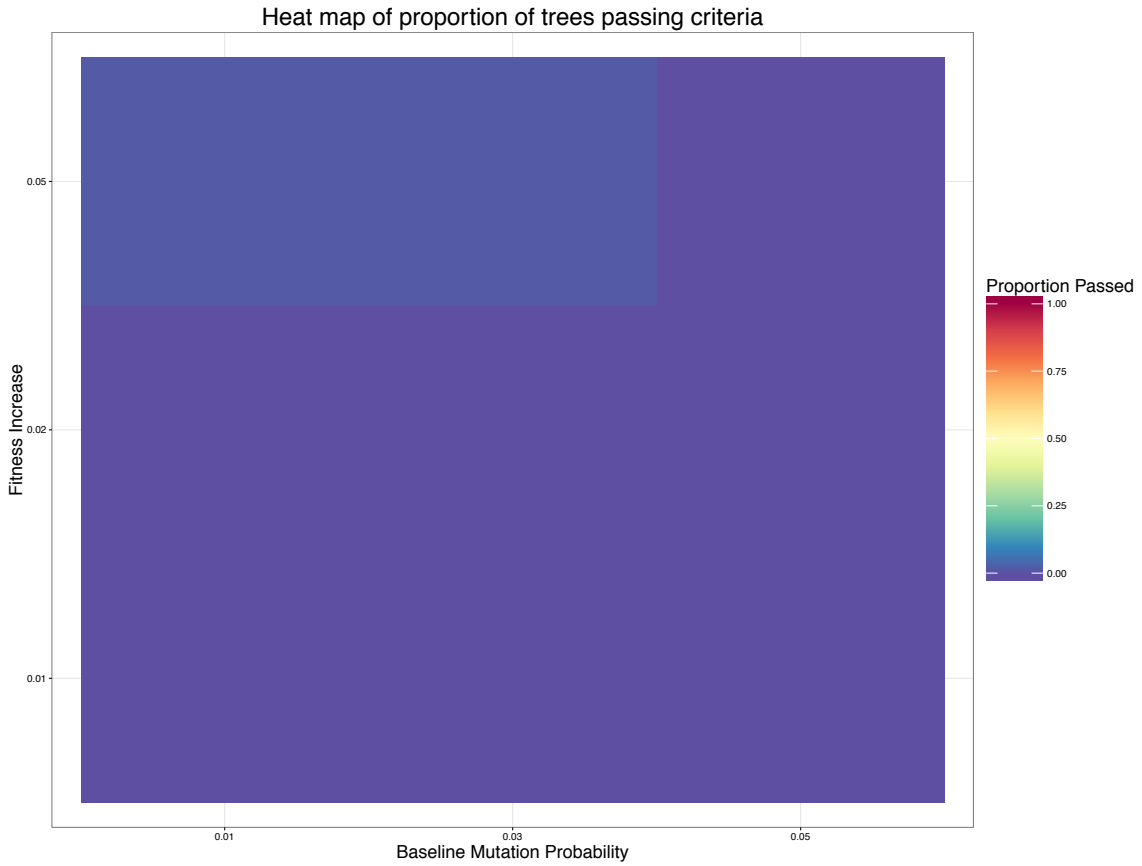
$$f(x) = \frac{1}{2}\alpha e^{-\alpha x} I(x \geq 0) + \frac{1}{2}\alpha e^{\alpha x} I(x < 0)$$




**S1.** The fitness distribution for a cell with a new CNA is chosen from a double exponential distribution in addition to the birthrate of the cells parent. The rate of the distribution will lead to differences in how fast populations grow and accumulate further mutations.

The maximum birth rate of any clone is capped at 1. We also used a fitness distribution of an atom at a single value so that the fitness increased by a fixed amount to determine if the shape of the fitness distribution has an effect on the population and trees created, and whether punctuated evolution could occur with fixed increases in fitness for subclones. The free parameters in this scenario include $\alpha$ - the rate parameter for the fitness distribution, and $u_0$ - the mutation rate.

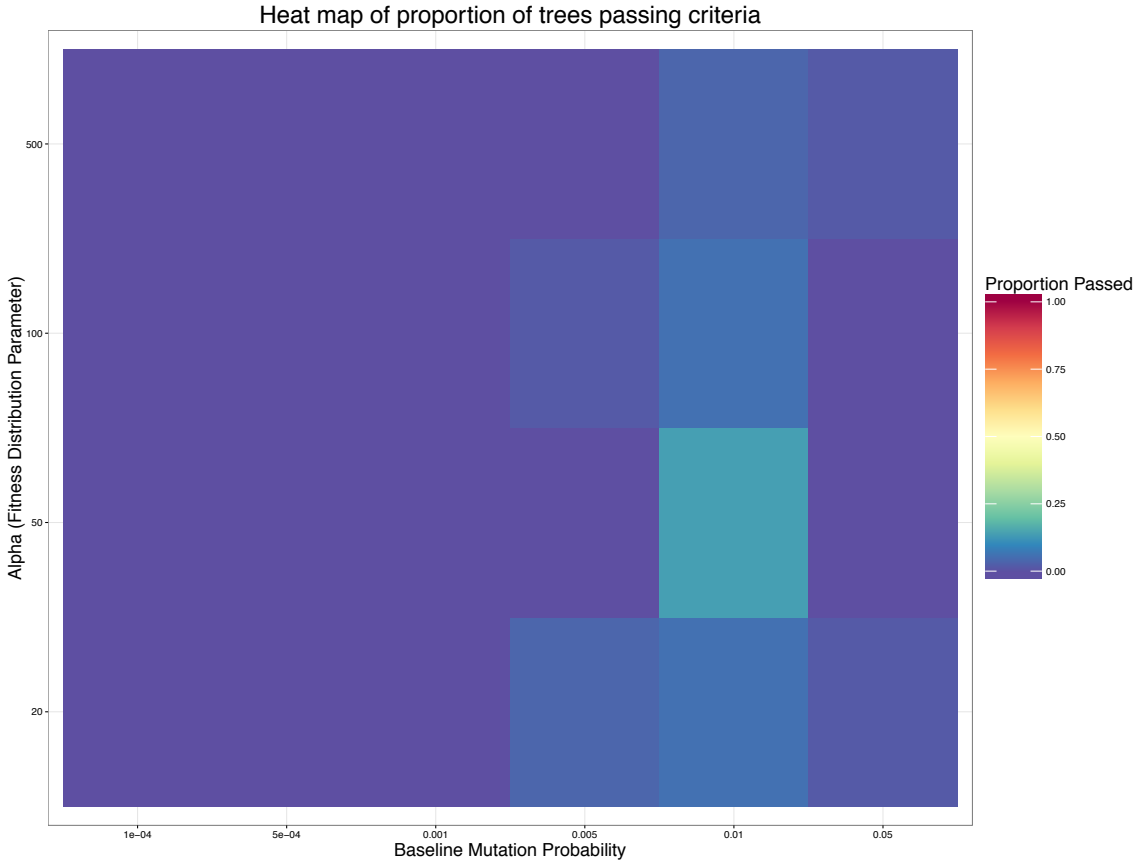
We also implemented a gradual model without using a fitness distribution to determine the fitness of mutant cells. Instead, we assumed a deterministic increase in fitness by specific amounts. If a parent has birth rate $b_0$ and gives rise to a mutant daughter, then the latter has birth rate $b_1 = b_0 + s$ where s is a constant value. We tested this model at fewer values of $s$ that cover a large interval in fitness. As new mutations arise, the increase in fitness should lead to subclones outcompeting parental clones by a fixed rate instead of choosing that rate from a distribution, guaranteeing waves of clonal expansions. The free parameter in this model is $s$ instead of $\alpha$.




**S2.** A model that uses a deterministic increase in fitness for gradual clonal evolution rarely shows passing scenarios based on different values for the amount of increase in fitness and the baseline mutation probability.

Figure S2 displays a heatmap with the proportion of simulations that passed at various values for the increase in fitness and the background mutation rate. The results of this heatmap show it is unlikely that a fixed increase in fitness would lead to trees that appear to have punctuated equilibrium, and are similar to results from the gradual model with random fitness increases described above.

We included a model considering the cancer stem cell hypothesis, that a small subset of cancer stem cells exist that undergo mutations and their continuous proliferation and diversification explain the punctuated effect. Under this hypothesis, we assume $10^{-3}$ CSCs per cell that may be an overestimation, and we reduce our population size to this. We assume the tumor bulk CN profiles are reflective of the profiles of CSCs since only CSCs have unlimited replicative potential, and variability accumulated in these cells can be propagated indefinitely and is reflected in the differentiated population. Figure S3 gives the results of this model also fail to reconstruct phylogenetic trees that contain a punctuated effect and pass our AMOVA tests.




**S3.** A heat map showing the results of testing the cancer stem cell hypothesis shows the model has poor performance with respect to AMOVA testing for punctuated trees. The model is the same as the gradual model with an end population at diagnosis of $10^{-3}$ times the population of all tumor cells at diagnosis.

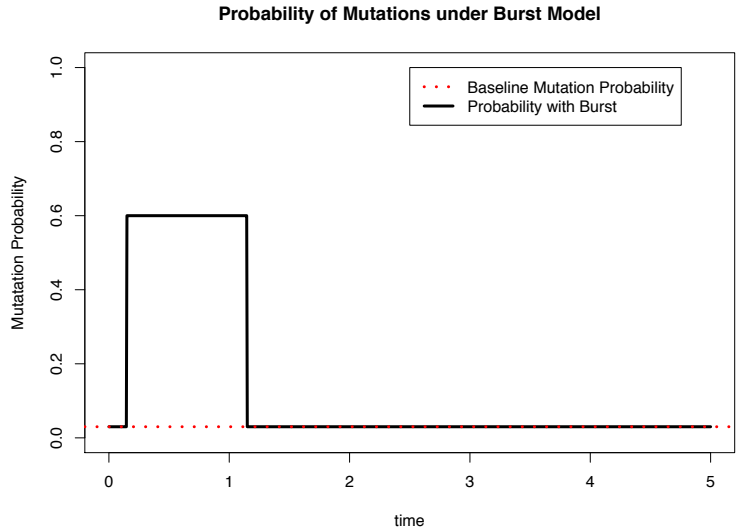
**2. The gradual model including varying mutation rates**

We next sought to investigate the effects of non-constant mutation rates. In order to consider variability between clones in the magnitude of their mutation rates, we implemented a variation of the base model, assuming a baseline fitness and mutation rate for the ancestor as before. Similarly, when a new mutation arises, the addition in birth rate is sampled from the same exponentially tailed p.d.f. as before. However, in contrast to the base model, we also included a distribution for the mutation rate such that a mutant daughter has mutation rate $u_0 + v$, where $v$ is exponentially decreasing. Values exceeding a mutation rate of 1 are reassigned 1 (i.e. any new cells have mutations as well). Since we assume a single CNA per mutant, this scenario could allow for a quick succession of clonal sweeps where subclones have higher mutation rates and fitness values than their parents and are able to initiate more fit subclones sufficiently quickly before the previous clone becomes too large. The free parameters in this model include $\alpha$ - the rate parameter for the fitness distribution, $u_0$ - the background mutation rate, and $\alpha_u$ - the rate parameter for the mutation distribution.

**3. The time-dependent mutation burst model**

We then aimed to study a scenario in which the mutation rate might significantly increase for a short time during the evolution of a tumor. This increase might be caused by environmental or

intrinsic factors and may contribute to the appearance of many CNAs during a short time in the evolutionary history of a tumor. The model assumes the same fitness distribution as model #1, and the mutation rate is either constant over time (as in model #1) or stems from a distribution (as in model #2). Additionally, in this scenario we consider a short-term increase (or "burst") in the mutation rate to occur early during the tumorigenetic process, where the probability of a mutation drastically increases in early clones before stabilizing according to a step function (Fig. S4). In this model, an early subclone can arise that has a larger mutation rate, thus generating further variability in its offspring at a rapid time scale. The free parameters in this model are $\alpha$ - the rate parameter for the fitness distribution, $u_0$- the background mutation rate, $\alpha_u$ - the rate parameter for the mutation distribution, $\tau$ - the length of time of a burst, and $\alpha_\tau$ - the rate parameter for the burst mutation distribution.

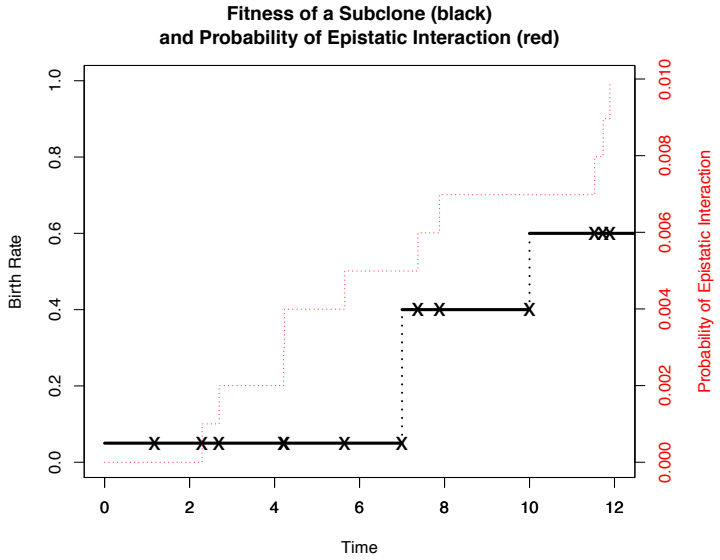



**S4**. Under the burst model, the probability of a CNA occurring during a time period is inflated such that early mutations have a larger chance of creating subclones with increased fitness and further mutations. After a period of time, the mutation rate stabilizes among cells back to the baseline mutation rate. The curve (black) shows the change in mutation rate compared to the model where mutation rate is constant (red).

**4. The epistasis model**

Next, we also considered a scenario in which the emergence of a CNA only provides a significant increase in the reproductive fitness of the cell on a specific genetic background, i.e. if sufficiently many CNAs have already been accumulated. The model assumes that CNAs affect fitness individually, but epistasis can occur which is considered to be an interaction of two or more CNAs that increases the additional fitness provided by each CNA by a given factor. That is, if a cell with k CNAs gives rise to a new mutant cell, there is a probability $(1 - p_{epi})^k$ that the new CNA does not interact with any previous CNAs. Thus epistasis occurs with probability $1 - (1 - p_{epi})^k$, and as a result the fitness distribution is altered by an increase in the rate parameter. This increases the probability of observing multiple CNAs in more fit clones, potentially leading to waves of clonal expansions. A sample trajectory of how fitness can increase over time in a clone along with the probability of epistatic interactions is shown in Fig. S5. The free parameters in this model are $\alpha$- the rate parameter for the fitness distribution, $u_0$ -

the background mutation rate, $\alpha_u$ - the rate parameter for the mutation distribution, $p_{epi}$ - the probability of epistatic interaction between any two CNAs, and $\alpha_{epi}$ - the rate parameter for the epistatic fitness distribution.

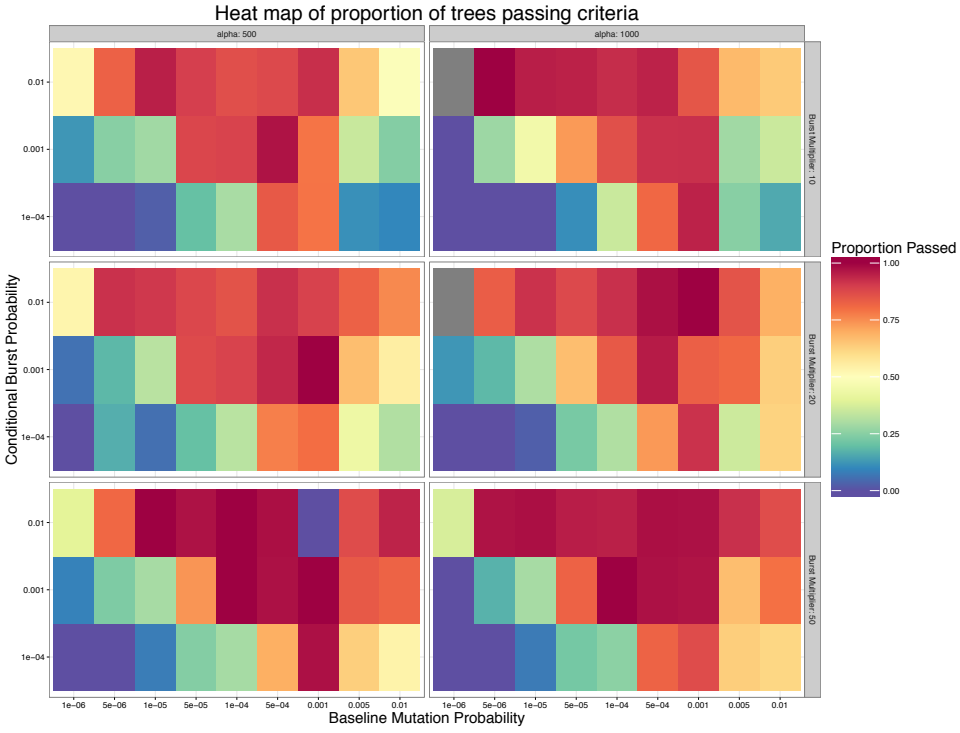



**S5.** As mutations accumulate, the probability of an epistatic interaction occurring (red curve) and leading to large changes in fitness (black curve) increases as well. Mutations (denotes by x's) in a single lineage's history are given as well as the affect on fitness (black) and the jumps in fitness from an epistatic event denoted by the gaps.

**5. The Poisson burst model**

Finally, we investigated a scenario in which multiple CNAs can be accumulated during a single cell division event. This model considers the same exponentially decreasing fitness distribution as model #1, but upon a mutation event occurring, the number of CNAs is chosen from a zero-truncated Poisson distribution with an inflated probability of observing a single CNA. We truncate at zero since the Poisson number of mutations is conditioned on a mutation event occurring. If an event occurs, then the event is either a single CNA with some probability $(1 - p_{burst})$ or a Poisson distributed number of CNAs with the complementary probability $(p_{burst})$. This scenario is based on the consideration of a standard baseline mutation rate where a single CNA occurs during any mutation event. At another, smaller rate, the event is more catastrophic, and multiple CNAs may arise during a single cell division. This model suggests that bursts of CNAs occur very rapidly, rather than by accumulating in successive sweeps. It also limits the probability of a second CNA arising out of the initial ancestor population before the tumor reaches detection size. We thus constrained the number of bursts to 1 since it seems likely that a second burst would be deleterious and would not appear in the tumor at the time of detection. Fig. S6 displays a heat map of the proportion of trees passing our criteria for the punctuated model while stratifying by different values for the rate of the fitness distribution for the initial clone and the increase in rate of the fitness distribution for the clone with Poisson mutations. These analyses uncover a tradeoff between the baseline mutation rate and the probability of a burst: a balance preventing an early clone from creating too many single CNAs while still providing a sufficiently large probability of a burst. The free parameters in this scenario include $\alpha$ - the rate parameter for the fitness distribution, $u_0$ - the mutation rate, and

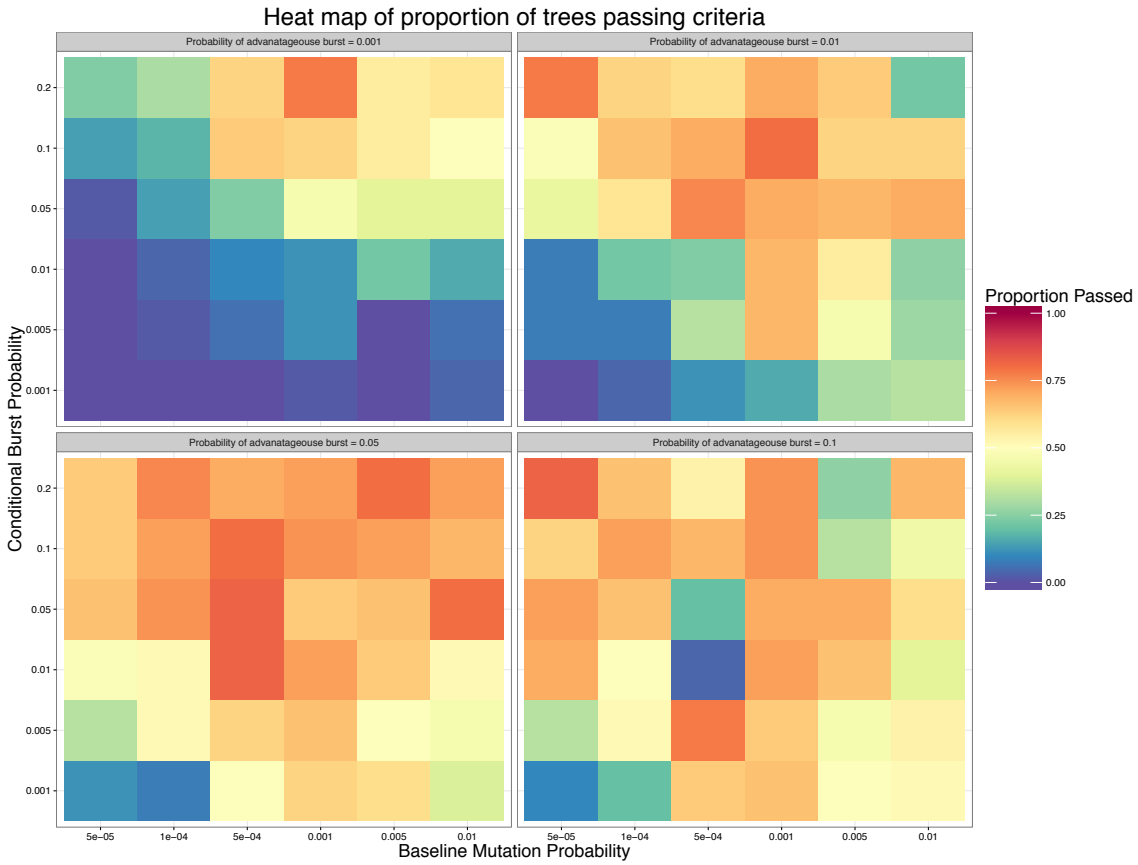
$p_{burst}$, the probability of a multiple mutation event occurring conditioned on a mutation event occurring. The parameter for lambda was fixed at 25, which is near the median difference in CNAs separating the diploid and the first aneuploid cell.




**S6.** Heat maps show much larger range of parameter values that pass our criteria for the punctuated model. The plots are faceted by the rate of the fitness distribution along the columns the increase in the rate of fitness distribution after a multiple copy number alteration change along the rows. Within each plot we compare across different values of the baseline mutation rate and the rate of a multiple copy number alteration event.

We also created a similar model including an extra step to account for the deleterious effect bursts of CNAs might have on the fitness of clones. In this scenario, there is an increase in the probability that a cell dies quickly, implemented as a fitness distribution of the death rate when a cell has a Poisson burst of CNAs. The model is the same as the Poisson burst model just described, apart from the fact that, when a burst occurs, the death rate jumps drastically with probability 1-q so that death is very likely. With probability q the death rate remains the same and we sample from our modified fitness distribution to describe the fitness advantage gained as above. In this model, the free parameters include: $u_0$ - the mutation rate, $p_{burst}$ - the probability of a multiple mutation event occurring conditioned on a mutation event occurring, and $q$ - the probability that a burst leads to a clone with a normal death rate instead of a high death rate that would occur with probability $1-q$. The parameter for $\lambda$ was fixed at 25, which is near the median difference in CNAs separating the diploid and the first aneuploid cell. We also set $\alpha$ in this scenario to 1000. Fig. S7 shows the results of testing simulations of the second burst model over a range of parameters. The number of simulations passed is still large, but there are a wider range of parameter values that do not reconstruct trees, which is most likely due to the lack of an advantageous burst occurring in the given time.

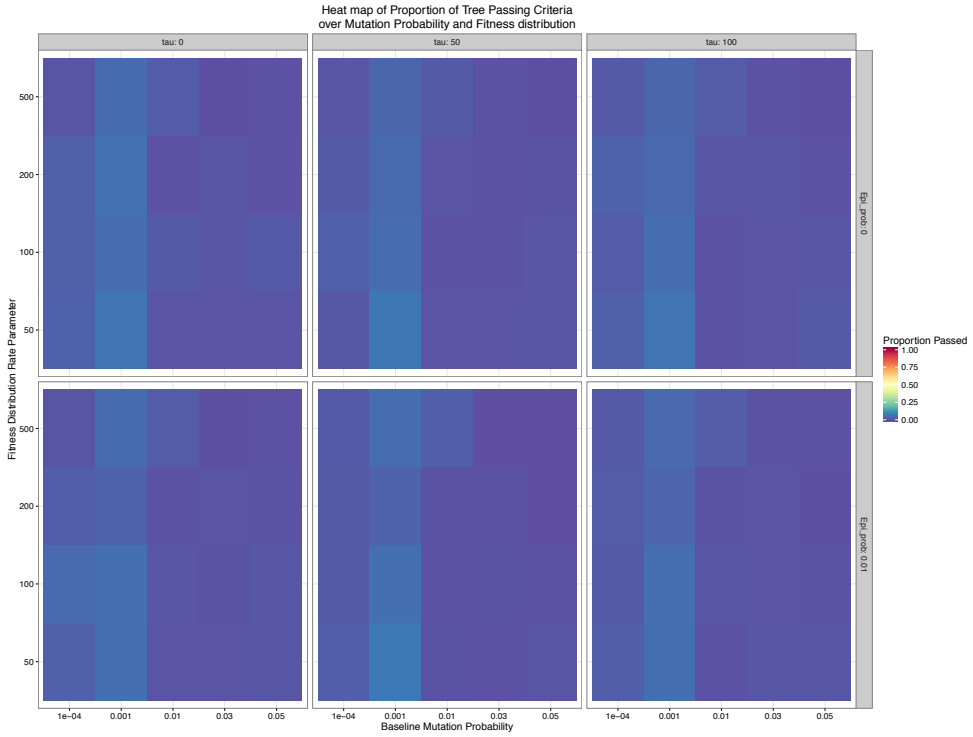
Figure S7 displays results from this new model of Poisson bursting with a high probability of death due to a burst. The model appears similar, and many simulations pass at a more specific combination of parameters. The results still show a great improvement over any gradual model of clonal evolution.




**S7.** A model with bursts of mutations results in a similarly large proportion of punctuated trees compared to the gradual model. The axes represent the baseline mutation probability and probability of a burst occurring conditional on a mutation event arising. The plot is faceted by the probability that the burst leads to a subclone with a fitness advantage as opposed to an increased death rate.

## Combination Models

To further investigate alternative scenarios, we also combined strategies from models 1 – 5 to determine if two different scenarios could occur simultaneously. The results from our comprehensive parameter and model search are summarized in Fig. S8, which displays a heat map from all combinations of the above strategies at low/medium/high values of the burst time and with and without epistasis. We found that models using most combinations of these assumptions performed poorly compared to model #5. Among all single CNA only cases, no scenario had more than 9 of the 50 simulations pass our criteria. Only model #5 had a large proportion of scenarios pass our criteria, where we allowed multiple CNAs to occur in single cell divisions.




**S8.** Heat maps show results from all scenarios in the gradual models when faceted by different burst times and with and without an epistatic probability term. In all scenarios, no set of parameters performed well. Rows denote different fitness distribution rates and columns are different baseline mutation probabilities. The color represents proportion of nonextinct trees that pass our criteria.

## Caveats and Discussion

We realize that models we created are not exhaustive, and represent only a subset of possible scenarios. We selected these models based on minimizing overfitting or creating an approach with too many rules. We acknowledge that the birth-death process does not consider competition for space or resources, as cells grow independently of each other. We also do not account for space in the model, and spatial models would be more realistic for solid tumors such as TNBC. Also, our ancestor is a diploid tumor cell, which we chose to initiate the process instead of starting with a population of normal breast cells. A process involving normal cells would require more stringent assumptions on the maximum size which changes drastically early in a person's lifetime, during puberty, and during childbirth, and therefore we believe that such a model has too many variables to estimate, many of which have not yet been determined. We instead chose to err on the side of parsimony and conditioned our process on the initialization of a tumor. Another issue for discussion is the assumption of infinite sites, or that each mutation event initiates a completely new CNA/set of CNAs. This assumption allows us to assume that CNAs are unique and prevents convergent evolution. Considering the large number of divisions that occur, the probability of a CNA needs to be small enough so that we do not observe more CNA events than sites for CNAs. Taken together, despite the caveats that apply to our model design and selection, our results overwhelmingly support the punctuated evolution hypothesis. We implemented a second Poisson model to investigate its robustness and consistency in recreating

punctuated trees. The second Poisson model with death parameters changing based on a burst more accurately describes the deleterious impact of genomic instability until a burst with a beneficial set of CNAs arises. The end result is still a model which predicts punctuation for certain sets of parameters, but those parameters are now shifted, and the expected time until the tumor reaches its diagnosis size is on average longer, reflecting the more commonly occurring cell death.